Audiovisual Integration of Reduced Information Speech Stimuli

A Senior Honors Thesis

Presented in Partial Fulfillment of the Requirements for graduation with distinction in Speech and Hearing Science in the undergraduate colleges of The Ohio State University

by

Meghan Hiss

The Ohio State University
June 2008

Project Advisor: Dr. Janet M. Weisenberger, Department of Speech and Hearing Science

Abstract

Every day, without knowing it, we are using more than one sense to perceive speech. Speech perception is a combined effort using not only auditory cues, but visual cues as well. This has been observed in situations where one of the cues is impaired, leading to a reliance on the other cue to fill in the missing pieces. An example of this would be a noisy environment where the auditory cue is difficult to interpret, and as a result, the individual will start to depend on his or her ability to interpret the visual cue. It has been found, however, that even when the auditory signal remains intact, individuals will still use their visual cues and fuse the two responses together. This is shown in the McGurk Effect, in which listeners were presented with an auditory stimulus of "ba" and a visual stimulus of "ga," with the result that most listeners perceived "da," a fusion of the two places of articulation.

Numerous additional studies have investigated the integration of auditory and visual cues in more detail. In general, three different aspects of the process have been identified as important determinants of audiovisual integration. Those aspects include talker characteristics, listener characteristics, and the effect of degrading the auditory stimulus. Previous studies in our lab have demonstrated the effects of degrading the auditory stimulus by reducing its spectral fine structure. Even with as few as four spectral channels of information, subjects have found these stimuli highly intelligible. However, another means of reducing spectral information in speech, a reduction to a series of three sine waves that follow the general formant structure of the stimulus, was found by our subjects to be far less intelligible. Because these previous studies employed different

groups of subjects, it is possible that observed differences in performance could be attributable to aspects other than the reduced waveforms themselves.

The present study addressed this question by performing a within-subjects comparison of intelligibility for these two types of auditory stimuli. In addition, we evaluated the potential priming effects that the order of the stimulus presentation had on performance for these two types of stimuli. 6 talkers and 12 listeners participated in this study. The 12 listeners were separated into three different groups, four participants to a group. The type of auditory stimulus and the order in which it was presented varied across groups. The two different stimuli used in this study were 2-filter degraded speech and sine wave speech. The stimuli were 8 CVC syllables, all of which had the same medial vowel and differed in only the initial consonant. The first group was presented the stimuli in an alternating order, i.e., the listeners listened to 2-filter degraded speech of a talker and then listened to sine wave speech of the same talker. The second group listened to all of the sine wave stimuli first and then listened to all the 2-filter degraded stimuli. The third group listened to all of the 2-filter degraded stimuli first and then listened to all of the sine wave stimuli. Each participant was tested under auditory-only presentation, followed by auditory plus visual presentation for each stimulus type. Results demonstrated that participants performed far better with 2-filter speech than sine wave speech. However, the order in which the stimulus was presented did not have a significant impact on the performance of the participants. Interestingly, subjects showed more audiovisual integration for sine wave speech than for the 2-filter speech, suggesting that a more highly degraded auditory stimulus promotes greater integration.

Acknowledgments

I would like to thank my advisor, Janet M. Weisenberger, for her constant support and guidance throughout the entire thesis process. Her dedication and experience has proven to be invaluable and I have grown professionally and academically as a result. I would like to thank Natalie Feleppelle for all of her time, encouragement, and support throughout this year. Furthermore, I would like to thank my subjects for their patience and flexibility while assisting me with my thesis.

This project was supported by an ASC Undergraduate Scholarship and an SBS Undergraduate Research Scholarship.

Table of Contents

Chapter 1: Introduction and Literature Review

Originally, speech perception was thought to be a process that relies on auditory cues alone, but further examination of this topic has suggested speech perception to be a multimodal process, that includes not only auditory cues, but visual cues as well. This multimodal process has been observed in situations where one of the cues is impaired, leading to a reliance on the other cue to fill in the missing pieces. When the auditory signal is compromised in some way, such as in a noisy environment or with individuals with hearing impairments, the result is that the individual will start to depend on his or her ability to interpret the visual cue. In these situations, the addition of visual cues can greatly improve an individual's ability to perceive speech. It has been found, however, that even when the auditory signal remains intact (i.e., completely intelligible), individuals will still use visual cues and fuse the two stimuli together.

Research by McGurk and McDonald (1976) demonstrated this fusion of the auditory and visual cues in what is known today as the McGurk Effect. McGurk and McDonald's study was conducted by pairing together conflicting auditory and visual cues, for example presenting the audio /ba/ with the visual /ga/, the auditory /pa/ with the visual /ka/, the auditory /ma/ with the visual /da/, and the auditory /va/ with the visual /da/ (Grant and Seitz, 1998). By pairing these phonemes together, McGurk and McDonald observed how participants integrated the auditory stimuli and the visual stimuli and attempted to determine if one mode would dictate the response. The results of McGurk and McDonald's study showed that when the auditory stimulus /ba/ was dubbed onto the visual stimulus /ga/, participants fused the two stimuli together and reported hearing /da/ (intermediate place of articulation). The results of this study indicate the occurrence of

integration between auditory and visual stimuli. McGurk and McDonald concluded that ignoring an input, in this case either the audio or visual input, is impractical in everyday life and that a person will employ all senses accessible to him or her in order to comprehend speech, even if only one modality is necessary.

McGurk and McDonald's study has sparked many questions regarding the integration process. One recurring question regards the conditions under which optimal integration occurs. Current research has focused on the aspects that prove to be critical components of the integration process. Researchers have posed questions regarding the circumstances that involve optimal speech integration: does the greatest amount of integration occur when the speech signal is highly intelligible, or when the speech signal has been compromised in some way? One thing we know for certain, in part from research from Shannon and his colleagues, is that the speech signal contains more information than is necessary to identify the sound.

Auditory Cues for Speech Perception

Speech waveforms contain acoustic spectral and temporal bandwidths that provide information regarding the place, manner, and voicing of a speech sound. Shannon *et al*. demonstrated the redundancy of speech signals in a 1995 study, where it was observed that even under conditions of reduced spectral information, high levels of speech recognition could be observed. In Shannon's study, spectral information was removed from speech by replacement of the frequency-specific information in a broad frequency region with a band-limited noise (Shannon *et al*., 1995). Thus, the temporal and amplitude cues were preserved in each spectral band, but the spectral detail within

each band was removed. Eight normal hearing listeners listened to 16 medial consonants, eight vowels, and simple sentences in each of the signal conditions. The results of the study showed that although speech recognition performance on all three measures increased with the number of noise bands, high speech recognition performance could be achieved with only three time-varying bands of noise representing the complex spectral pattern of speech. Shannon's study supported the notion that speech signals have a certain level of redundancy and thus that even when compromised, they can still be identified during speech recognition tasks. The results of this study sparked an interest in different signal degradation techniques.

Remez and his colleagues performed a similar study to Shannon; however, he reduced the acoustic waveform differently. Remez *et al*. (1981) used time-varying sinusoidal patterns that followed changing formant center frequencies of a naturally produced utterance as the stimuli for his study. This "sine wave" speech drastically reduces the information in the acoustic signal. The study consisted of three conditions in which independent groups of listeners were informed to different degrees about the tonal stimuli that they would hear. In the first condition, listeners were not informed about the sounds they would be hearing, but were asked to report their spontaneous impressions of the stimuli. The second group consisted of listeners who were informed about what they would be hearing and were asked to transcribe the utterances as accurately as possible. Remez *et al*. (1981) concluded from the first two instructional conditions that naïve listeners might not automatically perceive sinusoidal replicas of natural speech as linguistic entities, however, when instructed to do so, listeners performed well. The final group was asked to evaluate the speech quality of the tonal stimuli. This group was

informed that they would be presented with the sentence, “Where were you a year ago?” and were asked to make several judgments such as whether or not the sentence was intelligible and to provide a confidence rating for their judgments. Results of the study demonstrated a dependence on the level of prior knowledge of the stimulus presentation; individuals who were informed about the stimulus accurately described the utterances, even though a tremendous amount of information was missing from the speech signal, whereas individuals who had no prior knowledge regarding the stimuli were less accurate, but were still able to detect some linguistic content within the signal.

Visual Cues for Speech Perception

In the previously mentioned studies by McGurk and McDonald, Shannon and his colleagues, and Remez *et al*., the main focus was on the auditory signal and how manipulation of this signal can impact the intelligibility of speech. Those studies concentrated on auditory cues for identifying articulatory features such as place, manner, and voicing. Research has also been performed on the visual aspects of speech perception in an effort to identify the information provided by the visual cues. In contrast to auditory cues, which contain multiple articulatory components, visual cues primarily provide information regarding the place of articulation.

Because of the nature of the English language, problems can occur when relying on visual cues alone, especially in situations when certain sounds have similar visual characteristics. For example, the phonemes of /p/, /b/, and /m/ often cannot be distinguished from one another when simply relying on visual cues because they are all produced as bilabial consonants. The grouping of phonemes /p, b, m/ is known as a viseme grouping. Viseme grouping generally contain more than one phoneme. The term

viseme, which was coined by Fisher in 1968 (as cited by Jackson, 1988), refers to speech sounds that possess similar movement patterns. Visemes exist for both vowels and consonants (Jackson, 1988). One focal point of research has been the difference between vowels and consonants with respect to the visual characteristics each group possesses. A study performed by Binnie, Jackson, and Montgomery in 1976 (as cited in Jackson, 1988) using confusions of consonant-vowel syllables found the place of articulation to be the strongest perceptual feature in visual-only conditions. Furthermore, an inspection of various viseme systems revealed that the consonants /p, b, m/, /f, v/, and /θ/ are often grouped as visemes due to visible movements that are universal (Jackson, 1988). Vowels can also provide visual cues; each vowel is produced with a distinct oral cavity shape, meaning that no two vowels will look alike (Jackson, 1988).

It is important to note that visemes are not universal, due to differences in articulation among diverse talkers. The term homophenous, which was coined by Nitchie in 1930, refers to speech sounds or words that appeared alike on the lips and could not be distinguished by visual cues alone. In Nitchie's classification system (as cited in Jackson, 1988), consonants were grouped into homophenous, or visually identical, categories in which the within-category sounds differed from one another in voicing and/or nasality, but shared the same place of articulation. This term only applies to consonants because every English vowel is produced with a distinct oral cavity shape.

Auditory-Visual Integration Theories

The previously mentioned studies have all focused on the auditory and visual stimuli as separate processes. In this section we discuss the process of integration between the two and how this integration can lead to optimal speech perception. One

model commonly used to describe auditory-visual integration is the Fuzzy Logical Model of Perception (FLMP). The FLMP, according to Massaro (as cited in Grant, 2002), proposes that all sources of information (auditory, visual, and auditory-visual) are evaluated independently and that the information (cues) obtained from one source is compared to known descriptions in the memory to determine the degree to which the cues from a given source match alternative responses. All sources are integrated relative to prototypes (descriptions) in the memory to determine the overall degree of support needed for each response alternative (Massaro, 1987, as cited in Grant, 2002). According to Massaro (as cited in Grant, 2002), the multiplicative integration rule used in the FLMP is an optimal decision rule used to minimize the differences between obtained and predicted scores, and is therefore considered more of a fit to obtained bimodal scores rather than a prediction of optimal bimodal speech performance.

Grant (2002) notes that there are two reliable aspects of the FLMP that demonstrate this concept. First, the FLMP seeks to apply multiplicative integration to unimodal confusion data (i.e., the probability of responding *y* given *x*) to obtain a bimodal prediction, and second, human receivers frequently do better at identifying consonants than the FLMP predicts (Grant, 2002). Because the FLMP has been labeled a model promoting optimal integration, according to Massaro (1987) and Massaro and Cohen (2000), (as cited in Grant, 2002), this means that a poor performance in auditory-visual integration is the result of poor unimodal input, and not the result of poor integration abilities.

A different model that contrasts the FLMP is the prelabeling (PRE) model of integration. According to Braida (as cited in Grant, 2002), the PRE model does not seek

to optimally fit observed auditory-visual data, but seeks to "label" incoming bimodal stimuli based on an optimal combination of mutual information acquired from separate fits to auditory-only and visual-only performance. The PRE model obtains an estimate of unimodal information first and then, using an optimum combination rule, predicts how an unbiased receiver with no interference across modalities will do given the particular unimodal information accessible (Grant, 2002). In contrast to the FLMP, which promotes optimal integration, the PRE model accommodates the possibility that integration abilities might occur at a suboptimal level. Grant (2002) has stated that although data would indicate one model as being a better fit than the other, this does not necessarily imply that one model is more "correct" than the other, just that the two focus on different aspects of the integration process. However, for Grant and Seitz's 1998 study, they indicated the PRE model as being a better fit due to the model's ability to take into account individual differences seen in the speech perception of hearing-impaired individuals.

It is also important to keep in mind that auditory-visual integration proficiency is presumed to be a skill employed by subjects separately from their ability to extract information from auditory and speech inputs (Grant, 2002). Thus, auditory-visual integration is used to denote the *processes* employed by the individual receivers to combine information extracted from the separate auditory and visual stimuli, which is separate from the distinct ability to extract auditory and visual cues and the higher-order language processing of the information received by these two modalities (Massaro, 1998, as cited by Grant, 2002). Although research has supported the idea that the auditory signal (even when degraded) is highly redundant and contains more information than

necessary, there are still individuals who have difficulties identifying speech under these degraded circumstances. However, these individuals' ability to perceive speech often greatly improves under the condition of auditory plus visual stimulus presentation.

The two previously mentioned models, the FLMP and the PRE, were used to determine whether or not the ability to integrate auditory and visual cues was dependent on the integration efficiency each individual listener possessed. Grant and Seitz (1998) accounted for a range of potential individual differences in their study, such as degree of hearing loss, speechreading ability, and language skills. Having established that the integration process was independent of extracting the auditory and visual cues, Grant and Seitz (1998), were able to support the concept that not all hearing-impaired individuals will be able to speechread at the same level, due to the different levels of integration proficiency each person demonstrates. Research has shown that differences in individual integration efficiency result in a wide variety of performances during auditory-visual speech recognition tasks.

The Role of Auditory Information in Audiovisual Speech Integration

Auditory information is obviously a highly critical component of the auditory-visual speech integration process. As previously mentioned, the ability to understand speech can greatly improve with the addition of a second modality (i.e., the visual stimulus). But are there certain ways of degrading a speech stimulus that can render it almost impossible to perceive, even with the addition of the visual stimulus? Is the type of signal degradation the only factor determining the outcome of the integration process? Is it possible that previous exposure to one type of degraded auditory stimulus can affect performance with other types of stimulus reduction?

Previous studies in our lab have evaluated the impact of both sine wave and spectrally reduced speech stimuli on auditory-visual integration. In Huffman's study (2007), she focused on the characteristics of the auditory signal that would promote auditory-visual integration, specifically whether the removal of information from the signal would promote greater use of the visual input, leading to greater integration. The auditory stimuli used in Huffman's study were similar to the stimuli degraded by Shannon *et al*. (1995). Auditory syllables were reduced to a waveform composed of a broadband noise fine structure that was modulated by the temporal envelope of the original speech stimulus recording (Huffman, 2007). These degraded stimuli were then filtered into two, four, six, and eight spectral bands. The outcome of this filtering is the removal of the fine structure and discrete frequency information typically found in the speech signal, which effectively reduces the redundancy of the auditory stimulus (Huffman, 2007). The results showed that to some extent, listeners performed increasingly better auditorally when more spectral information was available; however, removing information from the auditory stimulus did not affect the degree of integration benefit (Huffman, 2007).

In contrast, another study performed in our lab by Tamosiunas used methods similar to Remez, but resulted in poor auditory-visual results for sine wave speech. His study addressed the question of whether reducing the redundancy in the auditory signal changes the auditory-visual integration process in either qualitative or quantitative ways (Tamosiunas, 2007). Tamosiunas investigated how auditory-visual integration occurs for isolated CVC syllables by presenting highly reduced, non-redundant speech cues in the form of sine waves together with visual speech information (2007). Under three

conditions: auditory-alone (A), visual-alone (V), and auditory plus visual (A+V), participants were asked to identify the sine wave speech syllables (Tamosiunas, 2007). Under the AV conditions, both congruent (Matching A and V phonemes) and discrepant (A phoneme is different from V phoneme) combinations were presented (Tamosiunas, 2007). Results showed that sine wave reduction of speech effectively reduces the available acoustic information found in the signal. This study also suggests that there may not be enough information contained in individual sine wave speech syllables to facilitate optimal auditory-visual integration. In fact, these dramatically reduced auditory stimuli actually impeded integration, because auditory plus visual performance was lower than visual-only performance across some sine wave arrangements (Tamosiunas, 2007). One possibility for the varying results between Huffman's and Tamosiunas' studies (2-filter speech being reported as intelligible and sine wave speech being reported as less intelligible) could be that these two studies employed different groups of subjects; therefore, it is not clear that the difference in results can be attributed solely to stimulus properties. The present study addresses these questions by performing a within-subjects comparison of intelligibility for these two types of auditory stimuli (2-filter and sine wave speech). If performance for the sine wave stimuli is found to be lower in this comparison, then it can be concluded that the sine wave stimuli are so reduced in information that audiovisual integration is not possible.

In addition, subgroups of subjects were tested with different order of stimulus presentation (e.g., sine wave speech followed by 2-filter speech, or vice versa), to assess whether exposure to one type of degraded speech influenced performance with another type of degraded speech.

Chapter 2: Method

Participants

In the present study, 12 participants were assigned the role of the listeners, while 6 other participants were assigned the task of being the talkers. The listeners consisted of 6 males and 6 females, ranging in age from 19 to 46. One listener reported some previous exposure to phonetics as part of her undergraduate major. The talkers consisted of 4 males and 2 females, ranging in age from 20-23, who produced a set of eight single syllable stimuli that were recorded by a video camera. All participants reported having normal hearing and normal or corrected vision. Eleven of the twelve listeners received ninety dollars payment for their participation, while one person was strictly a volunteer.

Interfaces for Stimulus Presentation

Degraded Auditory Signal Presentation

Each participant listened to two different types of degraded stimuli, one being a 2-filter degraded stimulus and the other being sine wave speech. Each participant sat inside a sound attenuating booth. During the presentation of the stimuli the participant received the auditory signal via TDH-39 circumaural headphones. Stimuli were presented at a comfortable suprathreshold level, approximately 75 dB SPL.

Degraded Auditory + Visual Signal Presentation

The presentation of the stimuli was the same in this condition, but in this case, a 50 cm video monitor located outside of the booth's double-glass windows (approximately

four feet away from the participants) was turned on to allow the participants to use both auditory and visual modalities in order to identify the stimuli.

Stimulus Selection

A limited set of eight CVC syllables was used as the stimuli for this study. Syllables were chosen according to the following specifications:

1. Pairs of the stimuli were minimal pairs, which differed in the initial consonant only.
2. All stimuli were accompanied by the vowel /æ/, which does not employ lip rounding or lip extension.
3. Multiple stimuli were used in each category of articulation, demonstrating place of articulation (bilabial and alveolar), manner of articulation (stop, fricative, and nasal), and voicing (voiced or unvoiced).
4. All stimuli were presented without a carrier phrase (citation style).

Stimuli

For each of the previously mentioned conditions, random sets of the same eight stimuli were presented. The closed set included:

1. Bat
2. Cat
3. Gat
4. Mat
5. Pat
6. Sat
7. Tat

8. Zat

Stimulus Presentation

Audio Signal Degrading

2-Channel Filtered Speech

For both types of auditory stimuli, the computer program Video Explosion Deluxe was used to record all six talkers. Each individual talker produced a set of eight CVC syllable stimuli a total of 5 times each. The stimuli were recorded through a microphone that was directly connected to a computer, which allowed the files to be stored in .wav format. For the 2-channel speech, the auditory files were input into a software subroutine created by Bertrand Delgutte for MATLAB. The auditory files were entered into the subroutine, which begins with two different stimuli, an input speech waveform and a broadband noise. The computer program then switched the amplitude envelope and fine structure of the two stimuli. The resulting stimulus composed of noise envelope and speech fine structure was discarded. From there, the remaining signal was filtered into two broad spectral bands. The cutoff frequencies for the two spectral bands were 80 Hz to 1877 Hz and 1877 Hz to 19.2 kHz. Similar to the stimuli of Shannon *et al*. (1998), the auditory files were reduced to a waveform consisting of a noise fine structure, which was modulated by the temporal envelope of the original stimulus recording.

3 Formant Sine Wave Speech

For the sine wave speech, the initial auditory files were input into Praat Version 4.4.29 computer software. A Praat script developed by Chris Darwin of The University of Sussex was used to reduce the auditory files to three sine waves centered on the first three formants (F0, F1, and F2). The auditory file (e.g., .wav) was converted to sine

waves based on the age and gender of the talker. The upper formant limits utilized were 5500 Hz for an adult female and 5000 Hz for an adult male.

Digital Video Editing

The visual stimuli in the current study were obtained from six talkers, four males and two females. A digital video camera was used to record the six talkers while each individual talker repeated the set of eight monosyllabic stimuli five times each. Once all the visual stimuli were acquired, the program Video Explosion Deluxe was used to edit both sets of stimuli (auditory and visual). The Video Explosion Deluxe software allowed for any auditory stimulus to be dubbed onto any visual stimulus. This allowed for the visual clips to be paired with not only normal auditory clips, but degraded auditory clips as well. The present study paired a talker's visual stimulus with that same talker's degraded auditory stimulus. A total of forty stimulus clips for each talker were produced using the Video Explosion Deluxe software. From there, the stimulus clips were input to a DVD burning software program called Sonic MY DVD, which allowed 60 stimuli to be burned onto a DVD. A total of thirty-six DVDs were used in the present study in two separate conditions, auditory only presentation and auditory plus visual presentation. Each individual listened to a total of twenty-four DVDs, 6 DVDs in auditory-only 2-filter, 6 DVDs in auditory-only sine wave speech, 6 DVDs in auditory plus visual 2-filter, and 6 DVDs in auditory plus visual sine wave speech.

Procedure

Testing Setup

Testing for this study was conducted in a basement lab room of Pressey Hall where The Ohio State University's Speech and Hearing Department is located. The lab contained sound-attenuating booths and digital equipment necessary for testing. Each subject was seated in a chair located against the back wall in the sound-attenuating booth. The chair was placed approximately 4 feet from the 50 cm video monitor, which was located outside of the booth's double-glass window. An intercom system was also located inside the booth and this allowed for communication between the subjects and the examiner.

Testing Presentation

Prior to testing, each listener was given a set of instructions to read. The instructions explained that under two different testing conditions (auditory only and auditory plus visual) talkers would be saying eight different words that only differed in the initial consonants. The listeners were informed that some of the words would be normal words, such as "mat" or "bat", but that other words would be words not in the English language, such as "zat" or "gat". The listeners were forewarned that the words they would hear would be degraded or "messed up" in some way. Listeners were instructed to concentrate on what they heard and if available, what they saw and to make a response after every stimulus presentation. The instructions stressed the importance of making a verbal response after each syllable presentation. There were approximately six seconds between presentations of each stimulus.

Testing Procedure

The 12 listeners who participated in this study were separated into three different groups, with each group containing four participants. The type of auditory stimulus and the order in which it was presented were the basis for the group separation. The two different stimuli used in this study were 2-filter degraded speech and sine wave speech. The first group was presented the stimuli in an alternating order, i.e., the listeners listened to 2-filter degraded speech of a talker and then listened to sine wave speech of the same talker. This was repeated until each of those four listeners had listened to all twenty-four DVDs. The second group listened to all of the sine wave stimuli first and then listened to all the 2-filter degraded stimuli. The third group listened to all of the 2-filter degraded stimuli first and then listened to all of the sine wave stimuli. The order of presentation within these constraints was the same for all listeners. In other words, each participant was tested under auditory-only presentation, followed by auditory plus visual presentation. Each trial was recorded by the examiner. Participants for the study devoted approximately three hours of their time. Rest periods for the participants were encouraged in order to minimize fatigue.

Chapter 3: Results and Discussion

Results for the two types of stimuli, 2-filter degraded speech and sine wave speech, were analyzed. Performance was evaluated for single-syllable (congruent) presentations for two different modalities, auditory only and auditory plus visual. The only difference, as previously mentioned, was the order of the stimulus presentation for each of the three groups. The results were scored based on percent correct performance. Auditory-visual integration can be assessed by comparing the performance between the degraded auditory only performance and the degraded auditory plus visual performance to see the increase in performance that results from an additional modality.

Percent Correct Performance

Figure 1 shows the overall percent correct performance averaged over twelve subjects for 2-filter degraded speech and sine wave speech by presentation condition (auditory only and auditory plus visual). There are several things worth noting from this figure. First, regardless of the degraded stimulus, whether it was 2-filter degraded speech or sine wave speech, auditory plus visual performance was better than auditory only performance. This suggests that the addition of a visual stimulus leads to an increase in performance, which supports previous studies that reported a benefit of two modalities rather than the use of only one.

A second note worth mentioning is that for both auditory-only and auditory plus visual conditions, the performance for the 2-filter stimuli was higher than the performance for the sine wave stimuli. This implies that there are significant differences in performance for both auditory only and auditory plus visual conditions between the

two types of stimuli, suggesting the use of sine wave speech results in poorer performance when compared to 2-channel filtered speech. Statistical analysis confirmed this observation, that significant differences in auditory only performance were observed, $t(11) = 9.538$, $p < .001$. In addition, significant differences in auditory plus visual performance were observed, $t(11) = 10.950$, $p < .001$.

Figures 2 and 3 indicate the effects of testing order. Figure 2 shows performance under auditory-only conditions for both types of stimuli (2-filter and sine wave speech) for all three testing orders (sine wave – 2 filter, 2 filter – sine wave, and intermixed). A 2-factor analysis of variance (ANOVA) evaluating stimulus type by testing order was performed to determine whether significant differences across testing orders were found. Not surprisingly, there was a significant main effect of stimulus type for auditory only performance, $F(1,9) = 102.729$, $p < .001$. However, for auditory only testing conditions there was no significant main effect of testing order, $F(2,9) = .620$, $p = .559$. Finally, for auditory only presentation, no significant interaction was found, $F(2,9) = 1.354$, $p = .306$.

Figure 3 shows the same breakdown as Figure 2, but for auditory plus visual performance. The patterns observed here are similar to those observed in Figure 2. For auditory plus visual performance there was a significant main effect of stimulus type, $F(1,9) = 91.456$, $p < .001$. There was, however, no main effect of the testing order, $F(2,9) = .731$, $p = .508$. Again, no significant interaction was found for auditory plus visual performance, $F(2,9) = .332$, $p = .726$. These results and the results for Figure 2 demonstrate the absence of a priming effect of testing order.

Figure 4 shows the percent improvement in performance between the auditory and auditory plus visual conditions for both types of stimuli. Improvement represents the

benefit that listeners are afforded by the addition of visual cues. Auditory-visual benefit is greater for sine wave speech. The increase in performance for the 2 channel filtered speech from auditory only to auditory plus visual was only 20% compared to that of sine wave speech, which had a benefit of 34%. The results suggest that speech stimuli degraded into sine waves benefits more from the addition of visual cues. This might indicate that auditory-visual benefit is enhanced when auditory stimuli contain significantly reduced amounts of information.

Overall, the 2 channel filtered stimuli were more intelligible than the sine wave stimuli. Regardless of the testing order, 2 channel filtered speech performance was consistently better across all test groups. Although there was no significant effect of testing order, it does appear that the group that listened to the sine wave stimuli first and then the 2 channel filtered stimuli had lower performance than the other groups. However, the degree of auditory-visual benefit was greater for the less intelligible auditory stimuli.

The results of the present study yield different findings compared to results of previous studies in our lab by Huffman and Tamosiunas. In Huffman's study (2007), the results for the 2-channel speech were 45% correct identification for auditory-only and 60% correct identification for auditory plus visual. In contrast, the results for the present study for the 2-channel speech were 66% correct identification for auditory-only and 85% correct for auditory plus visual. The difference between the two studies is over 20% for both modalities, auditory-only and auditory plus visual. These results could be attributed to a number of differences between the two studies, one being the employment of different 2-filter audio clips between the two studies. Another explanation for the

previous results could be that the two studies used different types of degraded stimuli, with Huffman using 2-filter, 4-filter, 6-filter, and 8-filter degraded stimuli and the present study using 2-filter degraded stimuli and sine wave speech.

In Tamosiunas' study (2007), the result for sine wave speech in auditory-only conditions was 13% correct identification. These results, are considerably lower than the present study's results, which yielded 30% correct identification of sine wave speech in auditory-only conditions. The difference between the results of these two studies could again be attributed to the different audio clips each study employed. Each study utilized different sine wave audio clips and there is the possibility that a difference in the quality of these clips affected the results between the two studies. Another explanation for the previous results could be that the two studies employed different sine wave reductions. Tamosiunas' study (2007) tested four different sine wave reductions: F0, F1, F2, and F0+F1+F2, whereas the present study tested 3 sine wave degraded speech stimuli. Therefore, it is possible that in Tamosiunas' study listeners became so frustrated with the unintelligible single sine wave stimuli that they became convinced that all of the audio conditions were unintelligible.

Chapter 4: Summary and Conclusion

Results of this study indicate that listeners performed better overall with the 2-filter degraded stimuli than with the sine wave stimuli. However, higher levels of audiovisual integration were observed with sine wave speech. This study also showed us that the order of the presentation stimulus does not have a significant effect on the performance of the individual listeners.

There were, however, some potential limitations to the present study. Because there were only four test subjects in each group, the results may not have revealed an effect of different orders of stimuli presentation. Replicating this study with more people per testing group may or may not yield different results.

Another potential limitation the present study may have possessed is the participants' lack of exposure to sine wave speech. All subjects who participated in this study were unfamiliar with sine wave speech and had never heard this type of stimulus. The same could be said for the 2-filter degraded speech, but because individuals already performed at higher levels than with the sine wave speech, longer exposure to 2-filter degraded speech might not yield any better results. Using subjects who are familiar with or who have had longer exposure to sine wave speech might have resulted in an increase in performance for that stimulus. In fact, Exner (2008) found that with longer exposure to sine wave speech, individuals performed at higher levels than when they were hearing the sine wave speech for the first time.

Results from this study suggest that the greatest integration between auditory and visual modalities occurs when the auditory signal is less intelligible. Overall, additional

studies are needed to further explore how the ambiguity of speech signal serves to facilitate or impede audiovisual integration.

References

Exner, M. (2008). Training Effects in Audio-Visual Integration of Sine Wave Speech *Senior Honors Thesis, the Ohio State University.*

Grant, K.W. (2002). Measures of auditory-visual integration for speech understanding: A theoretical perspective (L). *The Journal of the Acoustical Society of America, 112 (1)*, 30 – 33.

Grant, K.W. & Seitz, P.F. (1998). Measures of auditory-visual integration in nonsense syllables and sentences. *The Journal of the Acoustical Society of America, 104 (4)*, 2438 – 2449.

Green, K.P. (1998). The use of auditory and visual information during phonetic processing: implications for theories of speech perception. *Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-visual speech*, 3-25.

Huffman, C. (2007). The Role of Auditory Information in Audiovisual Speech Integration. *Senior Honors Thesis, the Ohio State University*.

Jackson, P.L. (1998). The theoretical minimal unit for visual speech perception: Visemes and coarticulation. *The Volta Review, 90 (5)*, 99 – 114.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746-748.

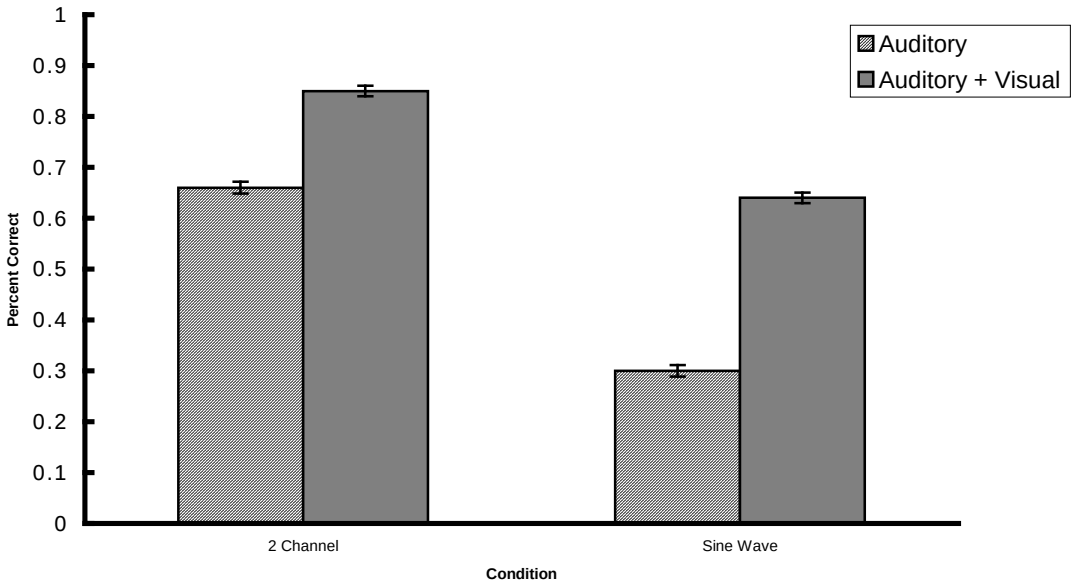
Remez, R.E., Rubin, P.E., Pisoni, D.B., & Carrell, T.D. (1981). Speech perception without traditional speech cues. *Science, 212 (4497)*, 947 – 949.

Shannon, R.V., Zeng, F.G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science, 270*, 303 – 304.

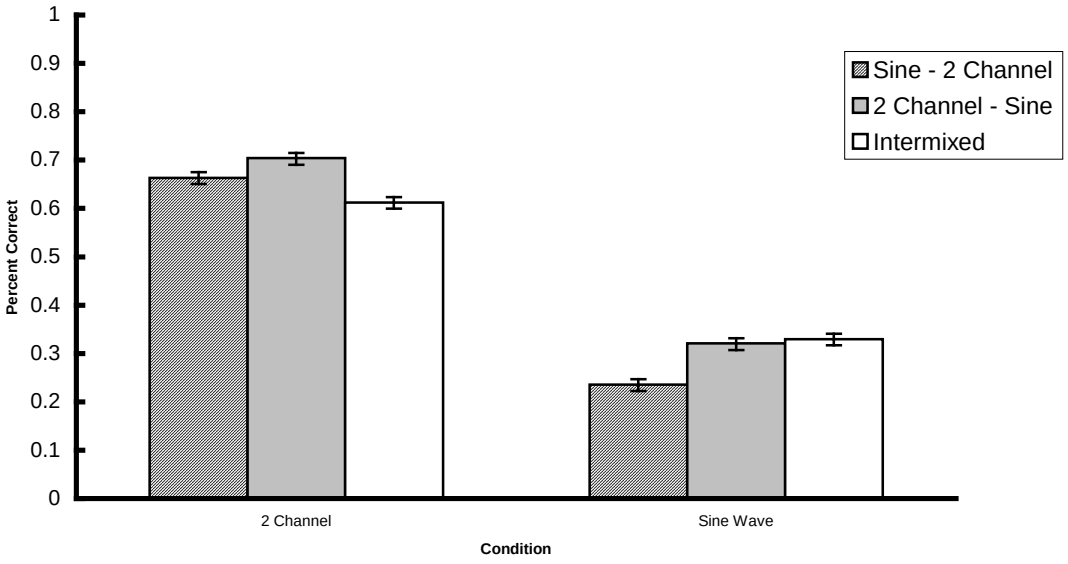
Tamosiunas, M. (2007). Auditory-Visual Integration of Sine-Wave Speech. *Senior Honors Thesis, the Ohio State University*.

List of Figures

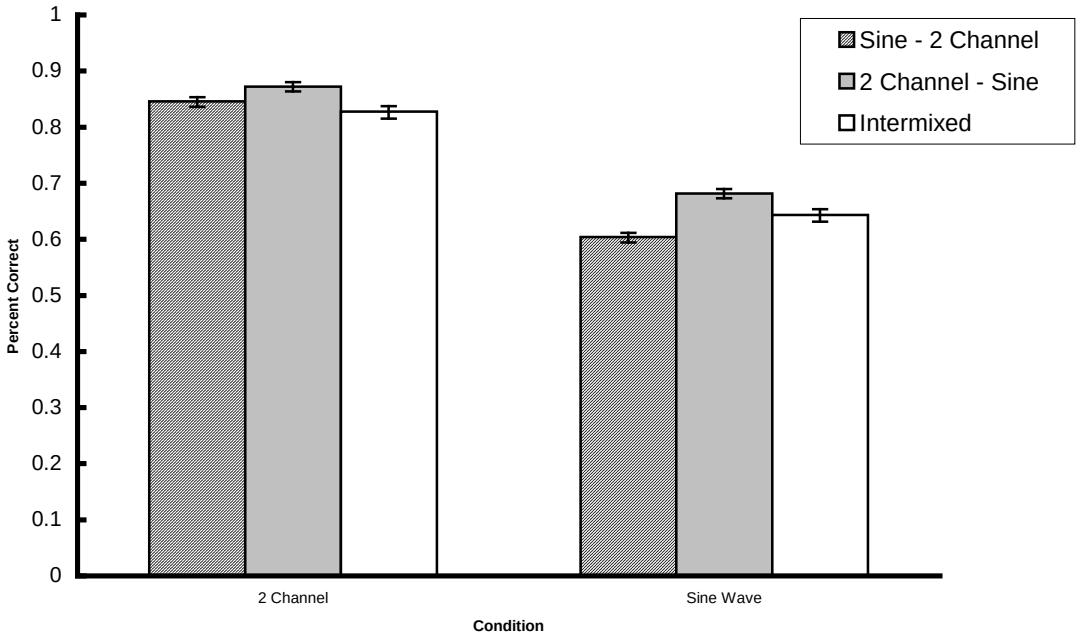
Overall Percent Correct By Testing Condition
Figure 1.



Auditory Only Performance by Testing Order
Figure 2.



Auditory Plus Visual Performance by Testing Order
Figure 3.



Percent Auditory-Visual Benefit by Test Material
Figure 4.